Reactive monologues

Modeling refinements and variations of interaction protocols of ECAs

Zsófia Ruttkay
Figures Bv, Amsterdam
zsofi@cwi.nl

Paul ten Hagen
Epictoid Bv.
paul.ten.hagen@epictoid.nl

Abstract At a recent Dagstuhl seminar on Evaluationg ECAs, the idea emerged of using the concept affordances and labeled graphs to describe interaction with ECAs on a high level. In this paper we explore how these two methods can be used to enrich information providing ECAs with perception capabilities, resulting in the illusion of personal and attentive services in contract to providing a straightforward and uniform dialogue.

1. Introduction

Embodied conversational agents (ECAs) [1] are models appearing on the computer screen, with the look and communicational modalities of real humans. There has been a wide range of ECA developed, different in the look, speech and multimodal communicational capabilities. They are gaining terrain as the user interface for or essential component of applications like information providing systems, e-commerce, tutoring, and entertainment.

Due to the novelty of the research field, the proliferation of modeling paradigms and implementation tools used, and last but not least, the inherent complexity of human-human communication, the design of the 'right' ECA for a given application is still more of art than science [7]. Another problem with present ECAs is that they are rather one-sided, as of communicating with the user. While researchers and designers are eager to endow their ECA with a multitude of subtle, human-like behaviors (talking with emotions, showing facial expressions, exhibiting elaborate eye-gaze patterns, accompanying speech with expressive hand gestures and postures), they pay less effort to the input side. Often the user is obliged to 'talk' to the ECA via the keyboard. If the user may provide input by speech, he easily gets frustrated by the shortcomings of the speech recognition system, limiting him in speech style and range of vocabulary. Recognition of (some characteristics of) the user, his emotional or cognitive state or even his presence are seldom thought of when designing the interaction capabilities of an ECA. This unbalanced approach can be understood in the light of the multitude of tasks to be taken care of on the synthesis side alone, and being aware of the fragility of the 'natural input' technology like speech understanding or facial expression recognition. The tasks on the recognition side are challenging themselves, and are dealt with in distinct research disciplines.

In spite of these arguments justifying, to some extent, the current practice in ECA research, one should look at an ECA as one of the parties in a scenario of interaction. An ECA may talk and gesture in a sophisticated way, but these capabilities will not compensate the clumsiness and inefficiency of communication due to lack of perception and reaction capabilities. You yourself may have experienced the strange effect of presenters at e.g. musea. You enter a room with all chairs empty in front of a screen. One possibility is that the lecture has already started, a talking head (videorecorded human or ECA) is already in the middle of his explanation. You may sit down for some time, but leave soon frustrated not knowing how long the entire talk will last (and afterwards, the missed part will start) … but the presenter keeps talking to the empty room. The other possibility is, when you sit down and wait patiently for the presenter to start, but he has not noticed you, and is waiting for some time-

scheduled starting. Such inattentive monolog is reminiscent of television and video, as media. However, ECAs as information providers usually do not perform much better, as of perceiving the user.

In our paper we will demonstrate how an ECA, initially presenting his monolog 'blindly', can be turned into an information provider attentive of and reactive to the user. As of 'how', we will address two aspects: the functional and the modeling aspect.

- From a functional point of view, we will identify ways how and why a monolog can be inappropriate, and suggest 'perceptive patches' to enhance the interactivity of the ECA.
- From a modeling point of view, we will use the unifying framework of labeled state transition graphs and the notion of affordances to point out the technical nature of refinements, and provide a basis for design and implementation.

The idea of such a modeling framework emerged at a recent Dagstuhl seminar dedicated to evaluating ECAs. A working group explored the benefit of the concept affordances to describe and design the interaction with ECAs [4]. The labeled graph framework was found to be promising, as making possible to design interaction with ECAs at the right level of abstraction and with the view on the desired goal of 'successful interaction. During the seminar, the initial ideas were tested on a few applications which had been developed by the participants. We got intrigued to investigate if the framework lends itself for designing *refinements* of interaction protocols. We chose the seemingly non-interactive application of information providers on purpose. We wish to demonstrate that even in the case of such a basically one-sided interaction there is reason and space for perceptive and reactive presentation. Also, we were eager to test and demonstrate the merit of the formal framework in a relatively single-sided (and thus, simple) communication protocol.

In the paper first we explain the notion of affordances in human-human interaction, and the usage of labeled graphs to model interaction. Then in section 3 we take an information provider ECA under the loop, and refine the interaction strategy from inattentive monologue to more attentive variants. In all cases, we present the refinements as modifications of the initial interaction graph. Finally, we discuss the merits for attentive monologues for some application domains, and raise further issues concerning the modeling framework.

2. A formal framework for designing interaction with ECAs

2.1 Affordances

The original notion *affordance*, invented by the perceptual psychologist J. J. Gibson [2, 3] was meant to refer to the actions which may be performed by a human (and also, animal) on objects of the world. He did not differentiate the action afforded (e.g. breaking a nut with a stone) and the physical, visual representation of the object as instrument for the action (the stone). With technological development, new physical devices have been invented to perform certain actions. E.g. nowadays when entering a room, one looks for the light-switch to turn the light on. While the basic function and mechanism of the switches are the same, there exist different designs. Besides the most usual ones are the up-down pegs, planar and turntable switches as well as touch switches exist. The visual appearance and the required usage are different, may be characteristic for the industrial design of a time period and/or country. People learn how to use a light switch as a child, by perception and repetition. (Turning lights on and off is one of the amusements of a 1-2 year old child.) Later on, the similarity in designs makes it possible that one recognises variants of designs, even new types, and can use them as intended. So, as Norman [6] pointed out, one should differentiate between the following aspects of affordances:

- *physical affordance*, standing for an object or device is capable to perform a certain task (to connect-disconnect electrical flow to the lamp);
- *visual representation* or appearance, which may be a range of variants for the very same physical affordance (different designs of switches);
- The (learnt) *capability of the human to identify the function and way of usage* from the visual appearance.

This refined concept of affordances has been proven fruitful in HCI for traditional user interface design, where one of the major tasks is to assign visual cues to possible actions on the screen.
Below we will explain how the concept can be used for the design of such a novel user interfaces as an ECA. First of all, we draw an analogy between affordances in human-object and human-human communication. When people communicate with each other, they use the 'physical affordances', such as articulating speech, showingfacial expressions etc. to convey information. What a major difference is between affordances in the world of acting with objects and in the world of acting between people is, that in human-human dialog the effect of an action (e.g. the effect of a sentence said) depends on the perceiving interlocutor too, who is a living person with a multitude of characteristics which may influence his reaction.

For instance, . the simplest sentence 'Give your bank account number', may remain without effect, because of very different reasons:

1. The partner is deaf, and thus the message did not reach him.
2. The partner does not understand English.
3. The partner did not understand well what was asked (due to intelligibility of speech).
4. The partner did not realise that he was addressed (in case there were more people around)..
5. The partner expects a more polite treatment, and does not 'obey' for such an authoritive command. Or simply he does not want to disclose the requested data, as he does not trust the partner enough.
6. The partner does not know how to give the answer (by filling in a form, or just typing in numbers, or talking to a microphone).
7. The partner cannot give the requested information, as he does not have a bank account.

Looking at the above cases from the point of view of affordances in human-human communication, the first two cases correspond to 'lack of physical affordance' on the listener side (of hearing speech, and of being able to process speech in English, as two condition for the entire physical affordance of English speech understanding). The third case corresponds to a 'malfunction' in using a physical affordance by the speaker (too low voice in a noisy environment, too fast speech for a non-native English speaker). The fourth case corresponds to 'not realizing the physical affordance being used', that is not realizing that the piece of speech was meant for the listening partner. Case five indicates that different aspects of the context of a single utterance (trust and power relationship between the conversants, the situation) influence the effect. Note that these aspects are judgements by the addressed person. In other words, there is no simple one to one relationship between 'action' and 'effect', as it is when acting with physical devices. Case 6 corresponds to the problem of not knowing which physical affordance is available for response, or how to use it. Finally, the last case is another example of the 'no effect' case, which depends on objective circumstances rather than the subjective skills and judgement of the addressee.

In real life situations, the cases like the ones listed above often occur, but usually do not halt the conversation. The problems in communication are intercepted, and recovered from, by:

- switching to an alternative physical affordance (in case of a deaf person, to written communication instead of speech, or trying another language than English),
- adjusting the usage of the physical affordance (slowing down the speech),
- making the adresse aware of the affordance (e.g. repeating the request with the name of the addressee, or establishing closer eye contact and coming/leaning closer, to make it clear that he is the one to answer),
- using a language more appropriate for the power relationship (as assumed by the addressee), or introducing a session to clarify the possible reason for delayed answer (lack of trust, lack of bank account, forgotten number).

Also in real life situations conversants try to minimize such mismatches in dialogues, by adjusting their communication to the changing parameters of the addressee and the physical circumstances, which are being more or less continuously perceived and processed. For instance, one makes a first guess of the language understanding and mental processing capabilities of the address by the look, which is corrected or refined in course of the conversation based on interpreting the verbal and nonverbal signals by the addressee. The speaker adjusts his speech and nonverbal modality parameters according to the changes in the physical parameters of the situation (with noise increasing, voice gets louder, more hand gestures may be used).

The aforementioned distinction of a physical affordance, its visual presentation to the user and the user's proper interpretation of the visual cue turns out to be useful to deal with, and design for, the complexity in ECA-human communication. The categories get the following interpretation then:

- A *physical affordance* is the capability to produce input or process it. E.g. the ECA is capable to process the speech by the user at certain moments of the communication, the physical affordance of natural speech (in itself assuming a lot of components like microphone on the computer, speech recognition system, NL understanding capability of the ECA) is available for the user.
- A *presentation of an affordance* is some indication to the interlocutor that at a given moment certain actions to produce input or feedback are possible or expected from him (e.g it is his turn to give an answer by speech).
- The *capability of the human* to identify the function and way of usage of the physical affordance, from the presentation. As argued above, this 'identification' may often go wrong, and because of very different reasons. Thus the success of a dialog largely depends on the monitoring of the dynamically changing parameters (of the user and in more general, of the environment), and adjusting the physical affordances and/or representation of them accordingly during the conversation.

2.2 Labeled state transition graph

Labeled state transition graphs are suitable formalism to capture all three aspects of affordances: input expected by some physical affordance at some point of the dialogue, the indication of availability of an affordance, and monitoring the success of exploiting the affordance. Transition graphs have been used to model interactive systems. In our case, specially, monitoring of and adjustment to the interlocutor can be explicitly modeled. In this paper we assume sequential processing. The speaker (which can be both the ECA and the human user) is in a state of one of the following types:

1. providing information (by using one or mode communicational modalities);
2. processing information;
3. monitoring the state of the interlocutor (and of the environment, which we will not deal with, for the time being).

There is a single state as starting state, and there may be one or more halting states. Labels on the arcs are inputs from the system (e.g text to be presented by the ECA), or result of the monitoring (e.g. perceiving if the user is listening or not). The flow of the dialog corresponds to some path in the graph. For successful interaction with ECAs, the monitoring states are of essential importance. As suggested by S. Marcella [5], the criteria of success of interaction should be also given in terms of characteristics of the covered path in the graph. This can be simply specifying some states which need to be included in the list of visited states (e.g. the user should be told about certain pieces of information), or the number of visits to certain states (e.g. keeping number of misunderstandings low in communication), or the duration of the interaction (e.g. for entertainment application, a long sessions are to be taken as success, while for travel booking shorter ones).

3. From monologue to attentive presentation

In this section we discuss a series of examples, each being *refinements* of a simpler talking head. By refinement we mean the refinement of the modeling labeled transition graph: a single node is replaced by a sub-graph. *Variants* are models with labeled graphs of basically identical structure, but with different labels.
For our discussion, we assume a museum information provider ECA, talking about paintings at a virtual Van Gogh exhibition. We assume that the ECA is a talking head, capable of speech, facial expressions and some hand gestures. We choose this, for the first sight straightforward and non-interactive application of an ECA because of two reasons:

- to demonstrate how much perception can enhance the effect of even basically 'monologues';
- to have a case simple enough to illustrate our points on affordances, and not to get overwhelmed with the problems of using several nonverbal communication modalities and a rich dialogue structure.

3.1 Start and stop talking

In the simplest form, the talking head gives a speech about some paintings of Van Gogh, similarly to a talk given by an expert for an audience. The only control available to the user is to start, suspend, continue or finish the presentation. In the graph of the interaction (see Fig. 1), the labels indicate the possible actions by the user. These actions by the user are perceived and processed, when the ECA is either in the initial idle state, or in the checking or waiting state. Arcs without labels indicate some routine cycles to a state checking if there is any input from the user, or that there is no action performed by the user in a state where user input is expected (idle, wait). These perception cycles are time-periodical. Hence, the speech by the ECA may get suspended in the middle of a sentence, or even word.

The physical affordances for the user to realize these actions can be, in the simplest form, buttons on the screen shown when applicable. The communication with the ECA can be made more human-like by eliminating the buttons as input devices, and making the ECA perceptive for the presence of the user. In such a case, the pushing of a start and finish button would be replaced by the perceiving if a use has arrived or has left. The affordances of the ECA for such a perception would be via vision, or by sensing pressure on the floor in front of the computer (see Figure 1.b).

3.2 Wrappings: greeting and introduction

In the previous scheme, the ECA reacts in an uniform way to all listeners and all the time. The refined presenter to be discussed in this section adds extra services. Similarly to a real lecturer, he greets his audience, and gives an overview about what he is going to present and reacts gracefully to the case when the listener quits. The modified state transition graph is given in Fig. 2.

There are some major improvements. If a listener appears in front of the ECA, some profile is made based on perceived characteristics. The richness of the profile depends on the perception capabilities of the ECA: is the ECA('s assotiated perception module) capable to derive the gender, ethnicity and age? Is it capable to recognize returning visitors? Based on the profile of the listener and some additional information (like time of the day), the ECA can greet the visitor in a more proper way than a 'Hello!'. He can say 'Good morning, madam.', or 'Hi, Mary, glad to see you again.' or 'Good afternoon, sir.' If ethnicity is recognized, the ECA would start talking in the language most suited for the ethnicity of the listener. Following the greeting, the ECA tells shortly what he is to talk about. Then he asks the listener if he is interested. If not, he says good-bye, otherwise he tells his story, checking from time to time if the listener is present. If not, he closes the session with taking farewell (which the listener will hear, as his absence is detected immediately.)

Note that the physical affordance to be used for making choice is not specified. The interest by the listener may have to be specified by selecting a yes/no button shown at the right moment, or in a natural way, if recognition of head movements and/or speech are supported. The dialog graph could be further refined with specifying how the ECA behaves while waiting for response from the user, and even repeating his question if no answer has been given for a long time.

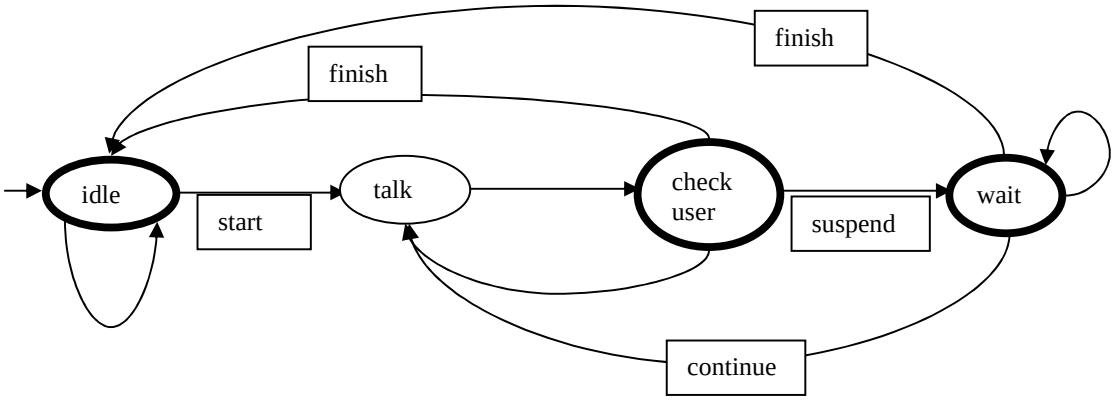

Fig. 1.a The simplest model of a talking head. Control is performed by some affordances (button, speech input) directly.

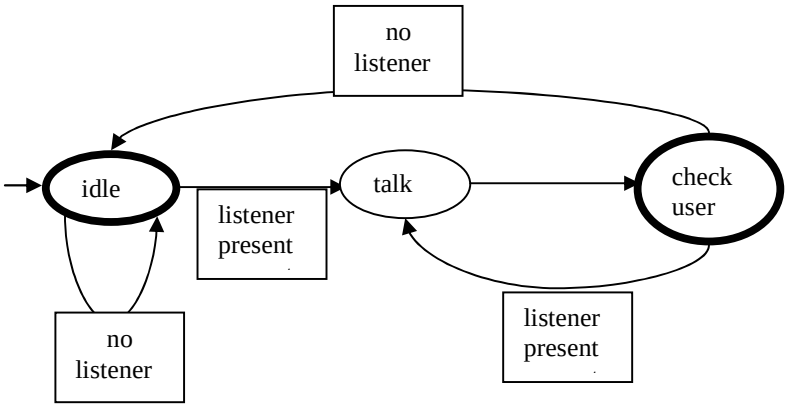

Fig. 1.b Part of the previous graph, but the control is performed on the basis of checking the presence of the listener.

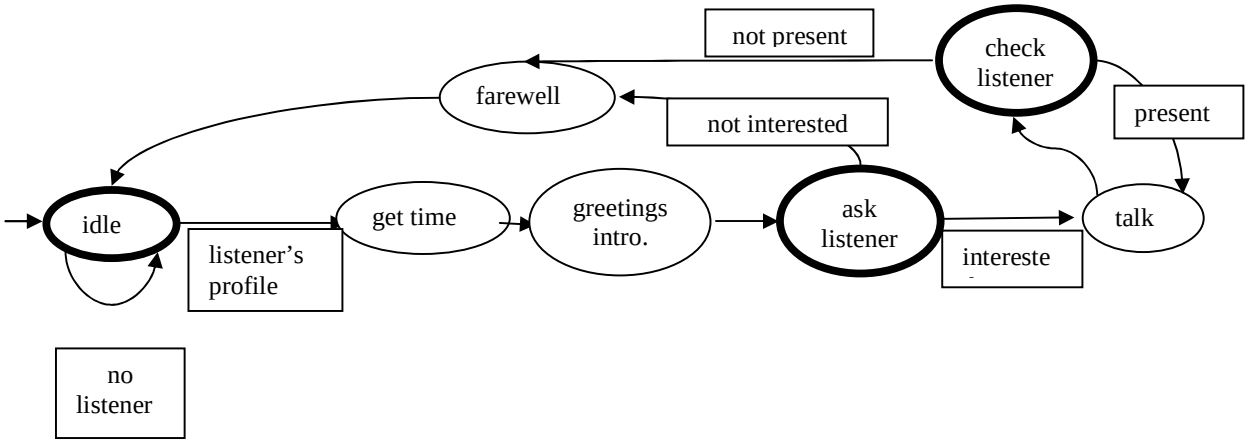

Fig. 2 The monologue shown in Fig 1 got extended with greeting and farewell referring tailored to perceived characteristics of the user and time of the day. Before starting the entire presentation, the listener has to decide to commit or not.

3.3 Topic and depth of information on demand

In the above schemes, each listener is given the same, possibly long talk. A monologue can be further improved by breaking the total chunk of information into smaller units around different topics (like the life of the painter, the story on the painting, the technique of the painting,…). Instead of a sequential, uniform presentation, the topic and amount of information given can be also adjusted according to the (directly or indirectly) expressed interest of the listener. The listener is asked to choose a topic from an offered set, by some input affordance like selecting from a menu, or telling his choice. While waiting for input (in the offer topics state), the ECA is indicating that it is the listener's turn to give an answer, e.g. by gazing at the listener. If no (valid) choice is made, the mechanism to make the selection is explained. Thus an affordance is offered to make the selection, and also, it is part of the dialog to draw the attention of the user to the input mechanism. This can be an explanation of how to use the input device (how to click, how to talk into the microphone).

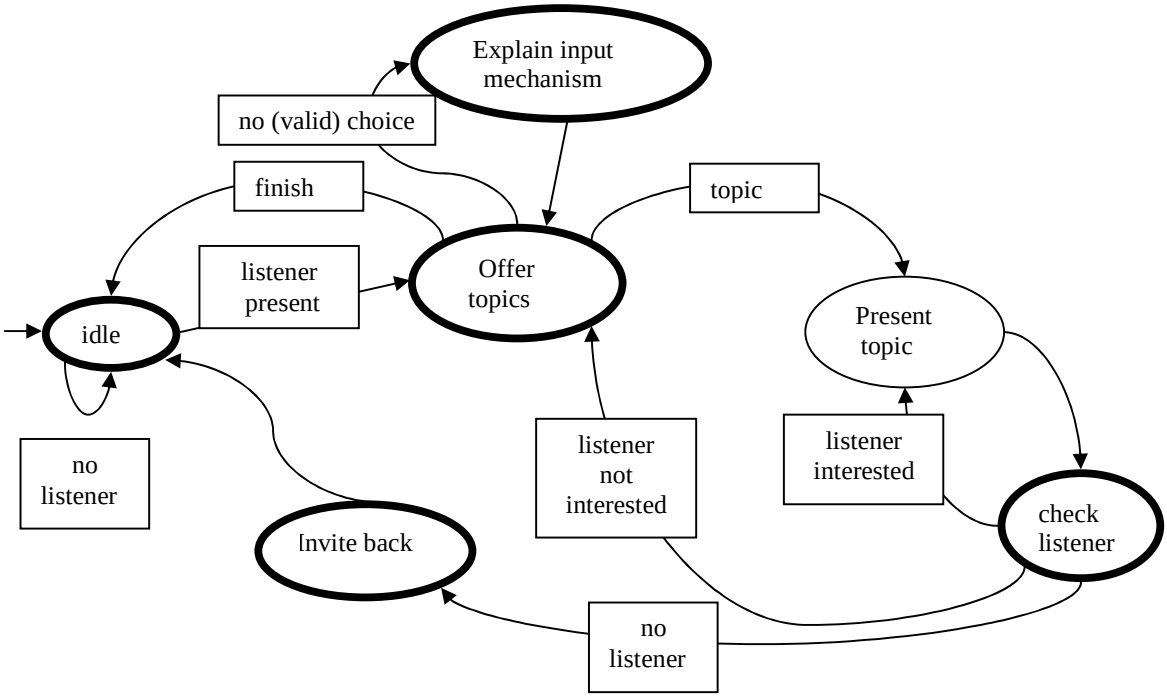

Fig. 3 Presentation with topic choice and explanation of input mechanism. There are actions for perception by the ECA and getting input from the listener by an affordance.

4. Summary

We are interested in extending the presentation capabilities of an ECA with perceptive and reactive capabilities. From a modeling point of view, we suggested a formalism suitable to model the input (perception), and output (presentation) stages as well as meta-information exchange on the control of interaction. We have argued that the concept of affordances is useful for interaction devices, both traditional ones as well as ones using human communication modalities, and labeled transition graphs are useful to model the flow of presentation, perception and input request actions by an ECA. We have also shown that perception of simple or more complex characteristics of the user (presence, facial expression) can be used to make the communication more person-tailored, and monitoring of and feedback on interest allows adjustment of the topics. These services can be seen as refinements of the initial, straightforward presentation by a talking head, which improve the quality. The freedom in

choice of physical affordances offered for the user, and the indication of them by the ECA (by using speech with non-verbal signals) allows realization of variants of the same dialogue scheme.

The consequent usage of state transitions and affordances allows a high-level conceptual design of interaction protocols for ECAs. Once the ideal (or rather, different LOD) interactions protocols are specified for ECAs in different roles, the functional affordances can be instantiated to physical input/output mechanisms, and the visual cues to be given by the ECA can be identified accordingly. To work in a sophisticated manner it would help very much if the total act of an ECA for a dialogue fragment could be obtained by merging the information contents acts with the affordances acts in a flexible, even dynamic, way. One can determine a set of expressive acts for ECA's that convey the dialogue control information interpreted as affordances. E.g. talking and listening (waiting for input) mode can be indicated by appropriate eye-gaze patterns, head movements and postures. ECA's allow for combining these acts parallel to or interwoven with acts that present contents (which, in a natural dialogue do not contain affordance information). These stereotypes can be used to create interaction schemes for applications. These can be translated into generic ECA performances in which what is actually to be said can be inserted to get a concrete peace of dialogue. The stereotypes are characterized by the sequence of affordances they convey.

Our examples for a presenter suggest that by using a few, cheap perception modules the illusion of attentive and personal conversation can be achieved. We believe that this would increase the acceptance and judgment of ECAs, especially in case of applications for users with different cultural and other characteristics.

In a next step, we wish to extend the model interaction by providing separate state transition graphs for the ECA and for the user, coupled by mutual perception of and feedback for each other's activity.

References

1. Cassell J., Sullivan J., Prevost S., Churchill E. *Embodied Conversational Agents*, MIT Press, Cambridge, MA. 2000.
2. Gibson, J. J. (1977). The theory of affordances. In R. E. Shaw & J. Bransford (Eds.), Perceiving, Acting, and Knowing. Hillsdale, NJ: Lawrence Erlbaum Associates.
3. Gibson, J. J. (1979). The Ecological Approach to Visual Perception. Boston: Houghton Mifflin.
4. Gratch, J. et al. Design criteria, techniques and case studies for creating and evaluating interactive experiences, Report of WG2 of the Dagstuhl seminar 04121 on Evaluating ECAs, 15-22 March, 2004. see http://www.dagstuhls/de/
5. Marsella, S. personal communication
6. Norman, D. A. Affordance, Conventions, and Design, INTERACTIONS, May 1999. pp 38-43.
7. Ruttkay, Z., Pelachaud, C. (eds). From Brows to Trust – Evaluating ECAs, Kluwer,to appear.